DSTO

AUTOMATIC SPEAKER RECOGNITION USING STATISTICAL MODELS

William Roberts

DSTO-RR-0131

APPROVED FOR PUBLIC RELEASE

DEPARTMENT OF DEFENCE

DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION

AUTOMATIC SPEAKER RECOGNITION USING STATISTICAL MODELS

William Roberts

Information Technology Division
Electronics and Surveillance Research Laboratory

DSTO–RR–0131

ABSTRACT

In this report we describe the automatic identification of speakers from their voices. This process has application in forensics and in voice actuated security systems. The implementation and theoretic underpinnings of a statistical based speaker recognition system are presented in addition to the performance of the system on standard speech corpora. In a speaker verification experiment, the system yielded an equal error rate of under 5% when identical microphones are used for testing and training.

DEPARTMENT OF DEFENCE

DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION

Published by

DSTO Electronics and Surveillance Research Laboratory
PO Box 1500
Salisbury, South Australia, Australia 5108

Telephone: (08) 8259 5555
Facsimile: (08) 8259 6567

AR No. 010-533
June, 1998

APPROVED FOR PUBLIC RELEASE

Automatic Speaker Recognition Using Statistical Models

EXECUTIVE SUMMARY

The performance and robustness of current speaker recognition technology is such that this technology may be operationalized for defence and civilian application. Applications include, e.g., voice actuated security systems, such as door locks, computer access controls, or telephone banking systems. In these applications, voice access may be more convenient than traditional methods requiring users to remember code words or identification numbers. Another application of speaker recognition is its use for forensic purposes.

The most popular speaker recognition techniques are those based upon statistical modelling. These techniques are reliable, well known, easy to implement and have substantial commercial and academic support. The model based approach has led to the development of commercial off-the-shelf dictation systems, e.g. Dragon Dictate. However, the source code for these systems is generally not available. Thus their modification to address problems of defence interest, apart from dictation and command recognition under favorable conditions, is frequently not possible.

This report describes the design and implementation of a speaker identification system suitable for use in real-world applications such as those mentioned above. The system has been designed using a combination of standard signal processing techniques found in the literature and other techniques developed here. The performance of the system has been measured on standard evaluation corpora. When compared to a similar signal processing system from the literature the system here had superior performance and reduced computational complexity.

Author

William Roberts
Information Technology Division

William Roberts obtained the B.E. (hons.), B.Sc. (hons.) and the Ph.D. degrees from the University of Adelaide in 1990, the University of Adelaide in 1992, and George Mason University, Fairfax, VA, in 1996, respectively. His areas of interest are information theory, detection theory, and statistical signal processing applied to speech. He began working at DSTO as a vacation student in 1988 and he is currently a Research Scientist in Information Technology Division. In 1998 he was awarded a fellowship from the Japanese Society for the Promotion of Science allowing him to undertake post-doctoral studies at the Tokyo Institute of Technology.

Contents

1 Introduction

Automatic speaker recognition refers to the process of recognizing speakers from their voices. This process may be performed by comparing the utterance from a speaker of unknown identity with templates or models of various speakers of interest. The degree of similarity between the models and the utterance is then used to make a decision.

Examples of applications for automatic speaker recognition systems include voice actuated security systems such as door locks, computer access controls, or telephone banking systems. In these applications voice access may be more convenient than traditional methods which may require users to remember code words or a personal identification number (PIN). Another application of speaker recognition is its use for forensic purposes [1].

The speaker recognition problem, referring to the general area of recognizing speakers from their voices, may be subdivided into smaller problems. Speaker *verification* is the problem of deciding if an utterance is from a particular speaker or not. Speaker *identification* is the problem of deciding who is speaking in a given utterance. Problems involving only a set of known speakers are *closed set* problems, whereas *open set* problems may involve speakers who have never been encountered previously and for whom no model exists. When the actual sequence of words that are spoken is known the problem is *text-dependent* and when it is unknown the problem is *text-independent.*

The speaker recognition techniques considered in this paper are those based on prescribing statistical models for the speakers of interest and training these models with training data. Other approaches are possible, e.g., techniques based on long-term-statistics [2, 3, 4], and neural network approaches [5, 6]. Long-term-statistics are extreme characterizations of the spectral characteristics and lack discrimination power [7]. Neural network techniques have achieved similar performance to statistical model-based systems [5] but their performance is affected strongly by network architecture and the quality and quantity of training data [5, 7, 8, 9]. In addition, neural training algorithms learn network node "weights" which are difficult to relate to the physical phenomena being modelled [8].

The most successful statistical model for speaker recognition is the Gaussian mixture model (GMM) which is related to the hidden Markov model (HMM) used extensively in commercial speech recognition systems. These statistical models are generally trained by parameter estimation using the maximum likelihood (ML) criterion. Once trained, classification may be performed using the maximum *a posteriori* (MAP) decision rule. The optimality of this technique is discussed in [10, 11]. The close relationship between GMMs and HMMs will allow speaker recognition systems developed using GMMs to take advantage of the considerable research efforts currently being undertaken by both academia and industry in HMM based speech recognition.

Generally GMMs are used to model a feature vector obtained from the speech time domain waveform. The most commonly used feature vector consists of the so-called cepstral coefficients whose use has dominated speech and speaker recognition over recent years. This feature vector provides high data reduction, obtains high performance, and is robust to certain channel effects. However it is obtained by a non-linear transformation and consequently is not robust to additive noise.

In this report we develop the theory and discuss the implementation and performance

of a speaker recognition system using Gaussian mixture modelling of cepstral coefficients. The system was tested using the evaluation data of the 1996 National Institute of Standards and Technology (NIST) Speaker Evaluation Workshop. The NIST workshop addresses the text-independent, open set, speaker verification problem and comprises specific testing and training conditions designed to exercise systems under real-world conditions. Using the NIST evaluation, the system was compared to similar GMM-based systems produced by Massachusetts Institute of Technology, Lincoln Laboratory (MIT LL). The system developed here had superior performance compared to the equivalent MIT LL system and had comparable performance to the MIT LL baseline system [12]. The system developed here had less computational complexity in that it used significantly less GMM states.

The remainder of the report is organized as follows. Section 2 presents the statistical signal processing theory underpining the speaker recognition system. Section 3 presents the implementation details of the system, the speech corpora used, and the results obtained. Section 4 presents the conclusions and possible avenues for further work.

2 Statistical Speaker Recognition Theory

Speaker recognition is an hypothesis testing problem. The optimal decision rule, in the sense of minimizing the probability of error, is the maximum *a posteriori* (MAP) decision rule [13]. The MAP rule decodes an acoustic utterance y, as the speaker $\hat{S}$, for which

$$\hat{S} = \arg\max_{S} p(S|y) \tag{1}$$

$$= \arg\max_{S} p(y|S)p(S) \tag{2}$$

where $p(S|y)$ is the *a posteriori* probability measure of the speaker S given the utterance, $p(S)$ is the *a priori* probability measure of the speaker, and $p(y|S)$ is the conditional probability measure of the utterance given the speaker.

Speaker *verification* constitutes a binary detection problem in which case the MAP rule may be represented as the likelihood ratio test. Thus the decision rule becomes

$$\frac{p(y|S_T)}{p(y|S_B)} \underset{\text{Background}}{\overset{\text{Target}}{\gtrless}} \gamma \tag{3}$$

where $p(y|S_T)$ represents the probability measure of the speaker of interest or *target* speaker, $p(y|S_B)$ represents the probability measure of the non-target or *background* speakers, and γ represents the decision threshold. There are two types of decision errors in the verification problem. The first is that of deciding that the utterance is a target speaker when it is really a background speaker. The second is deciding that it is a background speaker when it is a target speaker. These two errors are referred to as *false alarms* and *misses* respectively. For a given test, the probabilities of these two types of errors may be traded off against each other by varying the decision threshold γ. A plot of the false alarm probability versus either the probability of miss or detection is referred to a detection error tradeoff (DET) curve [12] or a receiver operating characteristic (ROC) curve [13], respectively. Frequently the performance at the point of equal false alarm and miss

probability is quoted to allow easier comparisons between tests. The error at this point is referred to as the equal error rate (EER).

The probability measures used in the above tests are not explicitly available. The tests may be implemented by estimating the probability measures $p(y|S)$ from training data, and using these estimates in the test as if they were the actual probability measures. This approach is referred to as the plug-in (PI) method. The estimation procedure is facilitated by prescribing parametric models to the probability measures. This has the advantage that only the limited parameter sets of the model need to be estimated from the training data. Parameter set estimation is generally accomplished using the maximum likelihood (ML) criterion. Other estimation criteria, such as the maximum mutual information (MMI) approach and the minimum discrimination information (MDI) approach, have been considered in the literature. These methods were compared in [14]. When, as is usually the case, the training sequence is considerably longer than the test sequence, ML parameter estimation combined with the PI approach can be shown to be asymptotically optimal in the minimum probability of error sense [10, 11].

The most common parametric probability model for speaker recognition is the GMM which as an example of the more general HMM used extensively in speech recognition systems. HMMs are also referred to as probabilistic functions of the Markov chains [15, 16] and as Markov sources [17, 18]. In the next section, we present some of the salient points of the GMM.

2.1 GMM models

The GMM consists of a finite number of states that are visited according to a state probability. When a particular state is visited, a random process is generated according to a probability measure that is associated with the state. This output random process is observable, but the actual state from which the process originated is not. Thus the state is considered hidden. A state sequence is generated as the process evolves with time. The GMM is completely specified by a parameter set consisting of the state probabilities and the parameters of the state probability measures.

We now present the standard assumptions of the GMM. For notational convenience, we suppress the conditioning of the parameter set λ of the GMM on the particular speaker, as all speakers may be treated equally. Let $y = \{y_t, t = 1, \ldots, T\}, y_t \in \Re^K$, be a sequence of vectors generated by an GMM. Let $s = \{s_t, t = 1, \ldots, T\}, s_t \in \{1, \ldots, M\}$, be the sequence of states that generated y. We can express the model $p(y|\lambda)$ as

$$p(y|\lambda) = \sum_{s \in \mathcal{S}} p(y|s, \lambda) p(s|\lambda) \tag{4}$$

where $\mathcal{S}$ is the set of all possible sequences of states, $p(y|s, \lambda)$ is the output probability of y given the state sequence s, and $p(s|\lambda)$ is the probability of the state sequence, s. The observation vectors $\{y_t\}$ are assumed to be independent of each other given the state sequence $\{s_t\}$ and the state sequence is assumed to be independent. Thus

$$p(y|s, \lambda) = \prod_{t=1}^{T} p(y_t|s_t, \lambda) \tag{5}$$

and hence

$$p(y|\lambda) = \sum_{s \in \mathcal{S}} \prod_{t=1}^{T} p(s_t) p(y_t|s_t, \lambda). \tag{6}$$

The state output probability density function $p(y_t|s_t, \lambda)$ is Gaussian thus

$$p(y_t|s_t, \lambda) = \frac{\exp -\frac{1}{2}(y_t - \mu_{s_t})^{\#} \boldsymbol{R}_{s_t}^{-1} (y_t - \mu_{s_t})}{(2\pi)^{K/2} \det \boldsymbol{R}_{s_t}^{1/2}} \tag{7}$$

where $\boldsymbol{R}_{s_t}$ and μ_{s_t} are the state dependent covariance matrix and mean respectively.

The parameter set λ of the GMM must be estimated from training data. A computationally efficient algorithm based upon the expectation-maximization (EM) algorithm [19] is available for the iterative ML estimate of the parameter set. This solution was first derived by Baum *et al* [15, 16], for the more general case of the ML estimate of the parameter set of an HMM. The solution for the GMM is considerably simpler and it appears in [20, Eqs. (5)-(7)].

An alternative approach using the segmental k-means algorithm may also be performed. In this case, the likelihood function is approximated by

$$\sum_{s \in \mathcal{S}} p(y, s|\lambda) \approx \max_{s} p(y, s|\lambda). \tag{8}$$

This approach can be shown to yield similar results under certain conditions [21, 22, 23]. This technique finds a single most likely sequence of states, s^*. The same re-estimation formulas of [20, Eqs. (5)-(7)] may now be used with $p(s_t|y_t)$ replaced by

$$p(s_t|y_t) = \delta_{s_t s_t^*} \tag{9}$$

$$= \begin{cases} 1 & s_t = s_t^* \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

The ML parameter estimation criterion is not the only possible criterion. Other parameter estimation criteria are sometimes used, e.g., maximum mutual information (MMI), maximum discriminant information, (MDI), or minimization of the empirical error rate, but their implementation is significantly more complicated than the ML approach.

Both GMMs and HMMs are widely accepted as reliable statistical models for speech signals and they have been successfully applied for speaker recognition, speech recognition, and speech enhancement [24]. The difference between the two models is that in the HMM, the state transitions are Markovian, whereas the state transitions of the GMM are not. In [25] it is reported that the Markovian state transitions of the HMM do not provide any extra performance for the speaker identification problem.

There are two main interpretations for the way in which they model speech signals. In the acoustic modelling interpretation [26], each state corresponds to a given configuration of the vocal tract. In the spoken language interpretation [27], each state represents a particular phoneme. HMMs and GMMs are also suitable for modelling noise sources [28, 29, 11].

Both GMMs and HMMs are frequently used to model feature vectors obtained from the speech waveform. The most popular feature vectors are those obtained via the cepstrum [27]. For speaker recognition, the cepstral vectors are modelled by GMMs with diagonal covariances. The next section describes the cepstrum.

2.2 Cepstral coefficients

Cepstral techniques have been applied successfully to speech signals since their introduction in the 1960s [27] and they now constitute the "standard" speech representation for speech and speaker recognition. Picone [30], in a survey of modern speech recognition systems, found that 21 out of 26 non-neural-network based speech recognition systems used some form of cepstral processing.

The cepstrum is an example of a homomorphic [27] signal processing technique and is defined as the inverse Fourier transform of the log of the power spectral density of the signal. Thus the cepstral coefficients $c(n)$ are given by

$$c(n) = \int_{-\pi}^{\pi} \log S(\omega) e^{j\omega n} \frac{d\omega}{2\pi}. \tag{11}$$

The cepstrum has the ability to deconvolve signals. If a signal u is assumed to have been obtained from passing an excitation signal w through a linear filter with impulse response g, then it may be described as follows

$$u = w \otimes g \tag{12}$$

where $\otimes$ denotes convolution. In the frequency domain this is represented as

$$U(\omega) = W(\omega)G(\omega) \tag{13}$$

where $U(\omega), W(\omega)$, and $G(\omega)$ are the Fourier transforms of u, w, and g respectively. If we take the logarithm of both sides

$$\log U(\omega) = \log W(\omega) + \log G(\omega). \tag{14}$$

Hence in the log frequency domain, the components due to the excitation and filter response are additive and hence are potentially easier to separate. It is this idea that provides a rationale for the cepstral analysis of speech signals. Considering the acoustic model of speech production, the cepstrum of the speech signal has a component due to the vocal tract and an additive component due to the excitation produced by the vocal cords. Thus conventional signal processing techniques may be used to separate these components.

In GMM-based recognition, the cepstral vectors are assumed to be Gaussian with non-zero means and diagonal covariances. The justification for the diagonal covariance assumption is provided in [31], where it is shown that, under certain regularity conditions on $S(\omega)$, for every two fixed positive integers k and l

$$\lim_{L\to\infty} \lim_{K\to\infty} K\mathrm{cov}(\hat{c}(k), \hat{c}(l)) = \delta_{kl} \tag{15}$$

where $\delta_{kl} = 1$ if $k = l$ and zero otherwise and $\hat{c}(k)$ and $\hat{c}(l)$ are the kth and lth empirical cepstral coefficients obtained from a length L smoothed periodogram estimate of $S(\omega)$. Thus the empirical cepstral coefficients, calculated by a smoothed periodogram, are asymptotically uncorrelated with a variance of $1/K$.

Generally, approximately 20 of these so called "first-order" cepstral coefficients are kept for speaker recognition purposes. Often the cepstral feature vector is appended with

delta-cepstral coefficients, which are approximations to the time derivatives of the cepstral coefficients and possibly delta-delta cepstra, which are approximations to the second time derivatives of the cepstral coefficients [32].

Estimation of the cepstral coefficients from data requires an estimate of the spectrum $S(\omega)$. This estimate may be obtained using non-parametric spectral estimation (e.g. periodogram or Blackman-Tukey) or by parametric estimation usually based on autoregressive (AR) modelling. In this case, a recursion is available to obtain the cepstral coefficients directly from the AR coefficients without explicitly calculating the spectrum [26, p 230].

2.3 Robustness of Speaker Recognition

In general, speech and speaker recognition systems work very well when trained and tested under similar conditions. However the systems are often not robust and very large degradations in performance occur when there is mismatch between training and test conditions [33]. Indeed, this lack of robustness is considered the major stumbling block for the widespread introduction of both speaker and speech recognition systems. There are two distinct causes for mismatch between training and testing conditions: additive noise and channel distortions. We do not consider the effects of additive noise in this report as there are many tutorials on this subject (see, e.g., [34, 35] and references therein). In this report we concentrate on channel effects and ways to combat them.

Channel distortions are induced by variations in the transmission path from the speaker to the recognition system input. A channel has numerous components including the acoustics of the room in which the utterance is made, the recording microphone, and the path from microphone to the recognition system, which in some applications, is via the inherently variable public telephone network. The effects on a signal due to channel variations are referred to as convolutional noise. Mismatch occurs when the convolutional noise is different in testing than it was during training. Mismatch due to differences in recording microphones can have a particularly large impact on performance. In [33] it was found that very large degradations in performance occurred when a system was tested and trained using different microphones. Room acoustics may be a factor as often training of the system is performed using data that is obtained in anechoic chambers. Unless testing is also performed in such environs, mismatch will occur. There are two main techniques used for channel compensation. These are cepstral mean subtraction (CMS) and RASTA filtering.

Cepstral Mean Subtraction CMS is a straightforward and easy to implement technique for combating channel effects. The simplest implementation of CMS involves estimating the mean of each cepstral coefficient over the entire utterance. The resulting cepstral mean vector is then subtracted from each cepstral vector [36, 37]. The removal of the mean in the cepstral domain corresponds to the removal of the long term spectral characteristics of the time domain signal. The assumption of CMS is that these long term spectral characteristics are due to channel effects. Techniques where only a short-term cepstral mean estimate is calculated and subtracted have also been proposed (see, e.g., [38].)

RASTA The relative spectral (RASTA) [39] technique is an example of a filtering technique designed to suppress constant or slowly varying signal characteristics. There

are many other examples of similar filtering techniques (see, e.g. [35] and the references therein.) In [39] the details of a particular filter are given that produced substantial improvements in the case of speech corrupted by convolutional noise.

In the next section, we describe an implementation of a speaker recognition system based on the theory described in this section.

3 Implementation

In this section we describe the implementation of a speaker recognition system based on Gaussian mixture modelling of cepstral coefficients. We discuss the calculation of the cepstral coefficients, important modelling details, and the speech corpora used to test the system.

3.1 Preprocessing

Cepstral coefficients for a given speech vector are estimated from a spectral estimate obtained using the window method [40]. Specifically the speech utterance is divided into frames of $K = 100$ samples each and a smoothed periodogram estimate of the power spectral density is obtained for each frame. For each frame, a $5K$ length "super-frame" is formed by concatenating the two K length frames either side of the frame to the original frame. The autocorrelation sequence is obtained by the inverse fast Fourier transform (FFT) of the magnitude squared of the FFT of the super-frame. The autocorrelation sequence is windowed by a Hanning window of length $K/3$. The windowed autocorrelation sequence is FFT'ed to form an estimate of the spectrum. See [40] for an analysis of this procedure for calculating an estimate of the power spectral density. The cepstral coefficients are obtained from the real part of the inverse FFT of the log spectrum. The 20 cepstral coefficients saved for each frame forms the feature vector that is modelled as an GMM with diagonal covariances.

3.2 Modelling

Both the background speakers and the individual target speakers are modelled by a GMM with $M = 20$ states. This figure is substantially less than the 2048 state model used in the MIT LL system [12] and represents a considerable computational savings. The Gaussian mixtures are non-zero mean with diagonal covariance matrices. Training of the GMM is accomplished using the segmental K-means algorithm, see section 2.1. The training algorithm is initialized with models obtained by a random clustering of the training data. In [25, 20] this simple initialization procedure demonstrated similar performance to more elaborate procedures based on phonetic clustering. The EM training iterations are terminated when the difference in likelihood between successive iterations was less than 1%.

3.3 Speech Corpus

The systems described above have been implemented and evaluated using the data and tests of the 1996 NIST Speaker Recognition Evaluation Workshop. This workshop addresses the speaker verification problem and consists of a speech corpus and a series of tests designed to evaluate the effects of utterance length, speaker sex, and microphone type on speaker recognition performance. More details of the evaluation may be found in [41, 12].

For the purposes of this study we considered a subset, consisting of male speakers only, of the full evaluation. The models were trained on 2 minutes of speech from 2 different microphones. The testing utterance length was nominally 30 seconds. Results for this test were split according to the microphone used during testing. The so-called "matched" results are those where the testing microphone was used during training and the "mis-matched" results are those for when the different microphones were used during testing and training. The background model was trained using speakers from the 1996 development corpus. There were no target speakers present in the background data.

3.4 Results and discussion

Fig. 1 shows the performance under matched and mis-matched microphone conditions.

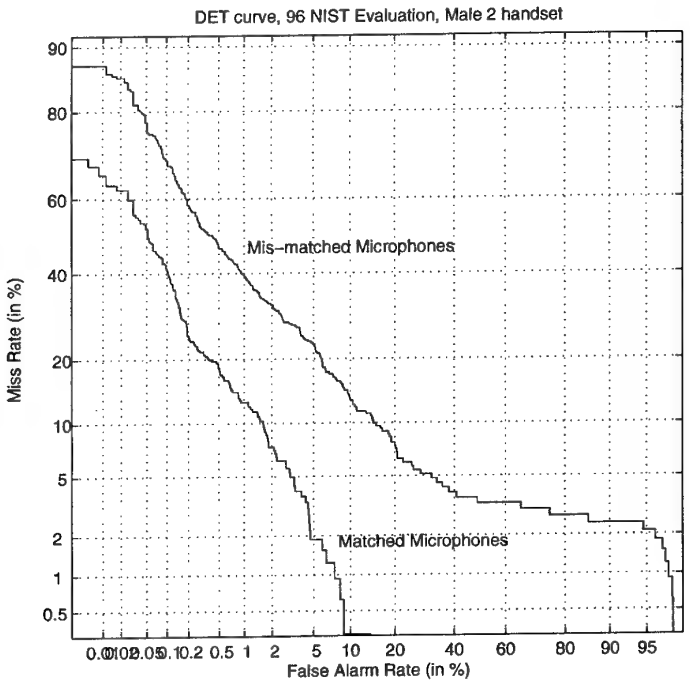

Figure 1: Performance of GMM system for matched and mis-matched microphones

The equal error rates (ERR) in Fig. 1 are approximately 5% and 14% under matched

and mis-matched conditions respectively. The MIT LL "UBM-independent" is a similar system to that developed here in that it used independently trained target models. When tested on the same subset of the NIST evaluation data as the system developed here, the UBM-independent system had matched and mismatched ERR rates of 8% and 20% respectively. The performance of MIT LL system was significantly improved by using a speaker adaptation technique, referred to as "UBM-adapt" in [12]. In this case the EER's were approximately 3% and 12%. The adaptation technique has the disadvantage that the so-called "relevance factor" [12, Eq. (7)] must generally be determined experimentally. Large increases in mis-matched performance were obtained in [12] via microphone adaptation. However this adaptation is difficult to implement in some applications.

4 Conclusion

The speaker recognition technology presented in this paper has a level of performance sufficient for it to be used in defence and civilian applications. The signal processing techniques used are well-known, easy to implement, and have demonstrated reliable performance by way of their use in commercial dictation systems. The system illustrated in this report provides performance comparable and in some cases superior to that of other systems in the literature. Further work in this area should address the relatively low performance under mis-matched microphone conditions.

Acknowledgements

The author appreciates technical discussion held with the following individuals: Doug Reynolds, Luc Gagnon, Richard Price, Jon Willmore, Lang White, Yariv Ephraim, Owen Kenny, and Lakshmi Narasimhan. The comments of the external reviewer, Robert Caprari, were also appreciated

References

1. G. R. Doddington, "Speaker recognition - Identifying people by their voices," *Proceedings of the IEEE*, vol. 27, no. 11, pp. 1651–1664, Nov. 1985.

2. S. Furui, "Talker recognition by longtime averaged speech spectrum," *Trans. IECE*, vol. 1, no. 55-A, pp. 549 -556, 1972.

3. J. D. Markel, B. T. Oshika, and A. H. Gray, "Long term feature averaging for sepaker recognition," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 25, pp. 330 337, 1977.

4. J. D. Markel and S. B. Davis, "Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 27, pp. 74–82, 1979.

5. Y. Bennani, F. Fogelman Soulie, and P. Gallinari, "A connectionist approach for automatic speaker identification," in *Conference Proc. IEEE ICASSP*, 1990, pp. 265–268.

6. J. Oglesby and J. S. Mason, "Optimization of neural models for speaker identification," in *Conference Proc. IEEE ICASSP*, 1990, pp. 261–264.

7. S. Furui, "An overview of speaker recognition technology," in *Advanced topics in automatic speech and speaker recognition*, C.-H. Less, Paliwal, and F. Soong, Eds. Kluwer, 1996.

8. L. Bruzzone, F. Roli, and S. B. Serpico, "Structured neural networks for signal classification," *Signal Processing*, vol. 64, no. 3, pp. 271–290, Feb. 1998.

9. J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the theory of neural computation*, Addison Wesley, Reading MA, 1991.

10. N. Merhav and Y Ephraim, "A Bayesian classification approach with application to speech recognition," *IEEE Trans. on Speech Processing*, vol. 39, pp. 2157–2166, Oct. 1991.

11. Y. Ephraim, "Gain-adapted hidden Markov models for recognition of clean and noisy speech," *IEEE Trans. on Speech Processing*, vol. 40, no. 6, pp. 1303–1316, June 1992.

12. D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *EUROSPEECH*, 1997.

13. H. L. Van Trees, *Detection, Estimation, and Modulation Theory*, vol. I, Wiley, New York, 1968.

14. Y. Ephraim, "On the relations between modeling approaches for speech recognition," *IEEE Trans. on Information Theory*, vol. 36, no. 2, pp. 372–380, Mar. 1990.

15. L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *Ann. Math. Statistics*, vol. 37, pp. 1554–1563, Dec. 1966.

16. L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," *Inequalities*, vol. 3, no. 1, pp. 1–8, 1972.

17. C. E. Shannon, "A mathematical theory of communication," *Bell Systems Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.

18. R. G. Gallager, *Information theory and reliable communication*, Wiley, N.Y., 1968.

19. A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm (with discussion)," *J. Royal Stat. Soc.*, vol. B39, pp. 1–38, 1977.

20. D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

21. N. Merhav and Y. Ephraim, "Maximum likelihood hidden Markov modeling using a dominant sequence of states," *IEEE Trans. on Speech Processing*, vol. 39, pp. 2111–2115, Sept. 1991.

22. N. Merhav and Y. Ephraim, "Hidden Markov modeling using a dominant state sequence with application to speech recognition," *Computer Speech and Language*, vol. 5, pp. 327–339, 1991.

23. Y. Ephraim, D. Malah, and B.-H. Juang, "On the application of HMMs for enhancing noisy speech," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 37, pp. 1846–1856, Dec. 1989.

24. Y. Ephraim, "Statistical model based speech enhancement systems," *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526–1555, Oct. 1992.

25. N. Z. Tishby, "On the application of mixture AR hidden Markov models to text independent speaker recognition," *IEEE Trans. on Speech Processing*, vol. 39, no. 3, pp. 563–570, Mar. 1991.

26. J. D. Markel and A. H. Gray Jr., *Linear Prediction of Speech*, Springer-Verlag, New York, 1976.

27. A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, Prentice Hall, Englewood Cliffs, NJ, 1989.

28. A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," in *Conference Proc. IEEE ICASSP*, 1990, pp. 845–848.

29. M. J. F. Gales and S. J. Young, "A fast and flexible implementation of parallel model combination," in *Conference Proc. IEEE ICASSP*, 1992, pp. 233–236.

30. J. W. Picone, "Signal modeling techniques in speech recognition," *Proceedings of the IEEE*, vol. 81, no. 9, pp. 1215–1247, Sept. 1993.

31. N. Merhav and C. H. Lee, "On the asymptotic statistical behavior of empirical cepstral coefficients," *IEEE Trans. on Speech Processing*, vol. 41, no. 5, pp. 1990–1993, May 1993.

32. B. A. Hanson and T. H. Applebaum, "Robust speaker-independent word recognition using static, dynamic and acceleration features: experiments with Lombard and noisy speech," in *Conference Proc. IEEE ICASSP*, 90, pp. 857–860.

33. A. Acero and R. M. Stern, "Environmental robustness in automatic speech recognition," in *Conference Proc. IEEE ICASSP*, Apr. 1990, pp. 849–852.

34. B.-H. Juang, "Speech recognition in adverse environments," *Computer Speech and Language*, vol. 5, pp. 275–294, 1991.

35. J.-C. Junqua and J.-P. Haton, *Robustness in automatic speech recognition*, Kluwer Academic Publishers, Norwell, M.A., 1996.

36. B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *Journal of the Acoustical Society of America*, vol. 55, pp. 1304–1312, June 1974.

37. S. Furui, "Automatic speaker verification," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 29, no. 2, pp. 254–272, Apr. 1981.

38. A. Rosenberg, C.-H. Lee, and F. Soong, "Cepstral channel normalization techniques for HMM-based speaker verification," in *International Conference on Spoken Language Processing*, 1994, pp. 1835–1838.

39. H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "Compensation for the effect of communication channel in auditory-like analysis of speech (RASTA-PLP)," in *EUROSPEECH*, 1988, pp. 1367–1370.

40. M. B. Priestely, *Spectral Analysis and Time Series*, Academic Press, 1992.

41. NIST, "March 1996 NIST speaker recognition workshop notebook," *NIST administered speaker recognition evaluation on the Switchboard corpus*, Mar. 1996.

Australian Speaker Recognition Using Statistical Models

William Roberts

(DSTO-RR-0131)

DISTRIBUTION LIST

Number of Copies

AUSTRALIA

DEFENCE ORGANISATION

S&T Program

Chief Defence Scientist	)	
FAS Science Policy	)	1 shared copy
AS Science Corporate Management	)	
Director General Science Policy Development		1
Counsellor, Defence Science, London		Doc Control Sheet
Counsellor, Defence Science, Washington		Doc Control Sheet
Scientific Adviser to MRDC Thailand		Doc Control Sheet
Director General Scientific Advisers and Trials	)	1 shared copy
Scientific Adviser - Policy and Command	)	
Navy Scientific Adviser		1 copy of Doc Control Sheet and 1 distribution list
Scientific Adviser - Army		Doc Control Sheet and 1 distribution list
Air Force Scientific Adviser		1
Director Trials		1

Aeronautical & Maritime Research Laboratory

Director	1

Electronics and Surveillance Research Laboratory

Director	1
Chief Information Technology Division	1
Research Leader Command & Control and Intelligence Systems	1
Research Leader Military Computing Systems	1
Research Leader Command, Control and Communications	1
Executive Officer, Information Technology Division	Doc Control Sheet
Head, Information Architectures Group	1
Head, Information Warfare Studies Group	Doc Control Sheet
Head, Software Systems Engineering Group	Doc Control Sheet
Head, Year 2000 Project	Doc Control Sheet
Head, Trusted Computer Systems Group	Doc Control Sheet
Head, Advanced Computer Capabilities Group	Doc Control Sheet
Head, Computer Systems Architecture Group	Doc Control Sheet
Head, Systems Simulation and Assessment Group	Doc Control Sheet
Head, Intelligence Systems Group	Doc Control Sheet

Head, CCIS Interoperability Lab	Doc Control Sheet
Head Command Support Systems Group	1
Head, C3I Operational Analysis Group	Doc Control Sheet
Head Information Management and Fusion Group	1
Head, Human Systems Integration Group	Doc Control Sheet
Head, C2 Australian Theatre	1
Task Manager	1
Author	1
Publications and Publicity Officer, ITD	1

DSTO Library and Archives

Library Fishermens Bend	1
Library Maribyrnong	1
Library Salisbury	2
Australian Archives	1
Library, MOD, Pyrmont	Doc Control Sheet

Capability Development Division

Director General Maritime Development	Doc Control Sheet
Director General Land Development	Doc Control Sheet
Director General C3I Development	Doc Control Sheet

Navy

SO (Science), Director of Naval Warfare, Maritime Headquarters Annex, Garden Island, NSW 2000.	Doc Control Sheet

Army

ABCA Office, G-1-34, Russell Offices, Canberra	4
SO (Science), DJFHQ(L), MILPO, Enoggera, Qld 4051	Doc Control Sheet
NAPOC QWG Engineer NBCD c/- DENGRS-A, HQ Engineer Centre Liverpool Military Area, NSW	Doc Control Sheet

Intelligence Program

DGSTA Defence Intelligence Organisation	1

Corporate Support Program (libraries)

OIC TRS Defence Regional Library, Canberra	1
Officer in Charge, Document Exchange Centre (DEC),	1
US Defence Technical Information Center,	2
UK Defence Research Information Centre,	2
Canada Defence Scientific Information Service,	1
NZ Defence Information Centre,	1
National Library of Australia,	1

Universities and Colleges

Australian Defence Force Academy	1
Library	1
Head of Aerospace and Mechanical Engineering	1
Deakin University, Serials Section (M list)), Deakin University Library, Geelong, 3217	1
Senior Librarian, Hargrave Library, Monash University	1

Librarian, Flinders University	1
Other Organisations	
NASA (Canberra)	1
AGPS	1
State Library of South Australia	1
Parliamentary Library, South Australia	1

OUTSIDE AUSTRALIA

Abstracting and Information Organisations	
INSPEC: Acquisitions Section Institution of Electrical Engineers	1
Library, Chemical Abstracts Reference Service	1
Engineering Societies Library, US	1
Materials Information, Cambridge Scientific Abstracts	1
Documents Librarian, The Center for Research Libraries, US	1
Information Exchange Agreement Partners	
Acquisitions Unit, Science Reference and Information Service, UK	1
Library - Exchange Desk, National Institute of Standards and Technology, US	1
SPARES	10
Total number of copies:	**64**

DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION
DOCUMENT CONTROL DATA

1. CAVEAT/PRIVACY MARKING

2. TITLE
Automatic Speaker Recognition Using Statistical Models

3. SECURITY CLASSIFICATION
Document (U)
Title (U)
Abstract (U)

4. AUTHOR(S)
William Roberts

5. CORPORATE AUTHOR
Electronics and Surveillance Research Laboratory
PO Box 1500
Salisbury, South Australia, Australia 5108

6a. DSTO NUMBER
DSTO–RR–0131

6b. AR NUMBER
010-533

6c. TYPE OF REPORT
Research Report

7. DOCUMENT DATE
June, 1998

8. FILE NUMBER

9. TASK NUMBER
DST 97/014

10. SPONSOR
DSTO

11. No OF PAGES
24

12. No OF REFS
41

13. DOWNGRADING / DELIMITING INSTRUCTIONS
Not Applicable

14. RELEASE AUTHORITY
Chief, Information Technology Division

15. SECONDARY RELEASE STATEMENT OF THIS DOCUMENT
Approved For Public Release

OVERSEAS ENQUIRIES OUTSIDE STATED LIMITATIONS SHOULD BE REFERRED THROUGH DOCUMENT EXCHANGE CENTRE, DIS NETWORK OFFICE, DEPT OF DEFENCE, CAMPBELL PARK OFFICES, CANBERRA, ACT 2600

16. DELIBERATE ANNOUNCEMENT
No Limitations

17. CITATION IN OTHER DOCUMENTS
No Limitations

18. DEFTEST DESCRIPTORS
Speech Recognition
Voice Data processing
Speech Communication
Voice Communication

19. ABSTRACT

In this report we describe the automatic identification of speakers from their voices. This process has application in forensics and in voice actuated security systems. The implementation and theoretic underpinnings of a statistical based speaker recognition system are presented in addition to the performance of the system on standard speech corpora. In a speaker verification experiment, the system yielded an equal error rate of under 5% when identical microphones are used for testing and training.